APPLICATION FOR PATENT

TITLE: SYSTEM TO OPTIMALLY ORDER CYCLES ORIGINATING FROM A SINGLE PHYSICAL LINK

INVENTORS: PARAS SHAH, RYAN HENSLEY and JAIDEEP DASTIDAR

SPECIFICATION

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0001] The invention relates to ordering cycles originating from multiple subordinate devices.

2. Description of the Related Art

[0002] Modern computer systems generally include input/output (I/O) devices that are connected to a central processing unit (CPU) via a system bus. The system bus operates to transfer addresses, data and control signals between the CPU and the I/O devices. Many modern computer systems include multiple buses, each in turn, with multiple I/O devices. Typically, any particular I/O device is coupled to only a single bus.

[0003] Bus bridges (bus-bridges) are often used in these multiple-bus systems to connect the multiple buses. In doing so, these bus-bridges receive communications to and from each of the multiple I/O devices connected to the multiple buses. "Bridge bridges" (bridge-bridges) are also often used in such systems to connect bus-bridges and thus handle communications from an even greater number of I/O devices. The commands transferred through both of these types of bridges frequently have data associated with them (e.g., read or write commands). The rate at which this multi-bridge architecture can process the communications generated from its multiple I/O devices directly affects the overall system performance. There is a constant demand for increasing the performance of modern computer systems generally. One way to achieve greater performance is to increase the rate at which communications from the I/O devices are processed.

[0004] As shown in FIG. 1, multi-bridge architectures can be viewed as having several different levels. A level 0 containing buses 120 and devices 110, a level 1 containing bus-bridges 130 and a level 2 including bridge-bridges 140. Level 0 includes I/O devices 110 connected to buses 120.

At level 1, bus-bridges 130 are connected to the buses 120 of level 0. Further, the bus bridges 130 have a transaction order queues (TOQ) 135 and transaction buffers 187 for each bus-bridge/bus link 160. TOQ 135 stores transaction buffer identifiers for certain transactions to ensure that system ordering rules, such as PCI and PCI-X ordering rules, are not violated. The purpose of the TOQ 135 is to ensure that transactions will execute in an order consistent with the system ordering rules. As such, not all transactions go into the TOQ, only those for which ordering rules apply. In contrast, transaction buffers store transaction information, such as cycle address, command, data, and the like. Next, level 2 contains bridge-bridges 140. Each bridge-bridge 140 is connected to one or more bus-bridges 130 from level 1. Each bridge-bridge/bus-bridge link 150 between bridge-bridge 140 and bus-bridge 150 has one representative TOQ 145 and one transaction buffer 185 in the corresponding bridge-bridge 140. An inherent difference between bus-bridges 130 and bridge-bridges 140 is that bus-bridges 130 connect a series of buses 120, while bridge-bridges 140 connect a series of bridges. The bus-bridges' 130 direct link to buses 120 assures that bus-bridges 130 always know from which one of the buses 120 a transaction originated. Bridge-bridges 140, in contrast, do not have a separate link for each of the buses 120, and as such, are not inherently able to identify the bus source of any transaction. This inability to identify the bus source negatively impacts the ability for ordering transactions at the bridge-bridge level, and as such, also unnecessarily limits the corresponding transactional throughput of the entire system.

[0005] Transaction ordering, as discussed in more detail in the two U.S. patent applications incorporated below: U.S. Patent Application, bearing Attorney Docket No. COMP:0189, entitled Relaxed Read Completion Ordering in a System Using a Transaction Order Queue, filed December 26, 2000, and U.S. Patent Application, bearing Attorney Docket No. COMP:0187, entitled Enhancement to Transaction Order Queue, filed February 8, 2001, orders a set of transactions based on a predefined set of rules. These rules are designed to achieve optimum transaction ordering where a single TOQ receives transactions originating from a single bus. However, where a TOQ receives transactions originating from multiple buses, optimum transaction ordering is lost and the overall transaction throughput is reduced. In further detail, and as shown in FIG. 1, TOQs 145 and 135, are used in two different types of bridges in two different levels. The first bridge, in level 2, is a bridge-bridge 140, where single TOQs 145 are used per each bridge-bridge/bus-bridge link (child-link) 150, regardless of the number of buses 120 attached to the corresponding bus-bridge 130. The second bridge, at level 1, is a bus-bridge 130, where single TOQs 135 are used for each bus-bridge/bus link (grandchild-link) 160. In the

case of a bus-bridge 130, where there exists a one-to-one ratio between TOQs 135 and buses 120, a TOQ 135, as designed, is limited to ordering the transactions from a single bus, and as such, is able to perform at its top design efficiency. However, in the case of a bridge-bridge 140, where there exists a one-to-many ratio between TOQs 145 to buses 120, a TOQ 140 is required to process transactions from multiple busses 120 over a single child-link 150. Specifically, TOQ 142 for example, is required to process transactions from multiple buses 122 and 124, and treat every transaction received through child-link 152, whether from bus 122 or bus 124, as though it originated from a single bus, and as such, the TOQ 142 is unable to function at its intended efficiency. In other words, because the bridge-bridge 140 is unable to discern between transactions of different buses 120 connected to a bus-bridge 130, the bridge-bridge 140 must order such transactions as though they occurred on the same bus, bus 122 for example. Because of this, unnecessary blocking occurs where a blocking condition on one bus, bus 122 for example, is imposed, across the entire child-link, child-link 152 for example, effecting every attached bus 120, and unnecessarily reduces transaction throughput.

SUMMARY OF THE INVENTION

[0006] Briefly, the illustrative system comprises a method and architecture for optimizing transaction ordering operations in a hierarchical bridge environment. The architecture includes at least a first bridge (parent-bridge), connected to a second bridge (child-bridge) via a link (child-link), and the child-bridge is connected to a transaction link (grandchild-link), where a parent-bridge has a set of buffers for each child-link to hold incoming transactions. For each child-link, the parent-bridge has at least two TOQs to provide separate transaction ordering for the child-links that communicate transactions from multiple different transaction sources, i.e., multiple grandchild-links.

[0007] In the illustrative system's most efficient operation, for every grandchild-link from the child-bridge, the parent-bridge maintains a dedicated TOQ. In providing a TOQ for each grandchild-link the parent-bridge is able to apply transaction order rules across the transactions from the individual grandchild-links in essentially the same manner as if the individual grandchild-links were each separately directly connected to the parent-bridge (i.e., child-links). This design essentially allows a parent-bridge to handle the transaction ordering of additional grandchild-links, without the need to connect such grandchild-links directly to the parent-bridge. Therefore, a parent-bridge with a fixed amount of child-links can virtually increase its number

of child-links by utilizing this multiple TOQ concept that essentially allows the parent-bridge to mimic the order processing that would take place if each separate grandchild-link had its own dedicated child-link to the parent-bridge. At a minimum, the illustrative system utilizes at least a two-to-one ratio of TOQs per child-link, and not less than a one-to-one ratio of TOQs per associated grandchild-link, and as such, is guaranteed to provide a higher level of transaction throughput than current one-to-one ratio TOQ-to-child-link systems.

[0008] Unlike the current systems that do not provide the means for a parent bridge to discern between the source of any communication received through child-link, the illustrative system provides such a means by transmitting an additional identification signal from the child-bridge to the parent-bridge. Thus, a signal can be passed from any child-bridge to its parent-bridge where the signal identifies from which particular grandchild-link a communication originated. Such a signal can be passed whether the parent-bridge is in level 2 and the child-bridge is in level 1, or parent-bridge is in level 3 and child-bridge is in level 2, or the parent-bridge is at any level "n" and child-bridge is at any level n-1.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] A better understanding of the present invention can be obtained when the following detailed description of the disclosed embodiment is considered in conjunction with the following drawings, in which:

Figure 1 is a component diagram showing a typical three level multi-tier bridge/bus/device architecture;

Figure 2 is a component diagram showing a three level multi-tier bridge/bus/device architecture utilizing multiple TOQs at level 2;

Figures 3A and 3B is a component diagram showing a four level multi-tier bridge/bus/device architecture utilizing multiple TOQs at multiple levels; and

Figure 4 is a block diagram showing the larger system for which an implementation like that of Figure 2 can be used.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENT

[0010] The following related patent application are hereby incorporated by reference as if set forth in its entirety:

[0011] U.S. Patent Application Serial No. 09/749,111, entitled "Relaxed Read Completion Ordering in a System Using a Transaction Order Queue," filed December 26, 2000; and U.S. Patent Application Serial No. 09/779,424, entitled "Enhancement to Transaction Order Queue," filed February 8, 2001.

[0012] Figure 2, illustrates a typical multi-bridge architecture 400 implemented according to the disclosed techniques. For purposes of explanation, specific embodiments are set forth to provide a thorough understanding of the present invention. However, it will be understood by one skilled in the art, from reading the disclosure, that the invention may be practiced without these details. Further, although the embodiments are described including three levels of bridges, most, if not all, aspects of the system illustrated apply to systems using two or more levels of bridges. Moreover, well-known elements, devices, process steps, and the like, and including, but not limited to, bridge and transaction order queue design, are not set forth in detail in order to avoid obscuring the disclosed system. As used herein, the "child-link" generally refers to the connection between the parent-bridge and child-bridge, including the link interface of the parent-bridge or the link interface of the child-bridge. Also, the term "grandchild-link" generally refers to the connection between the child-bridge and the grandchild component, including the link interface of the child-bridge or the link interface of the grandchild component. In Figure 2, the multi-tier bridge architecture 400 in the illustrated embodiment is made up of 3 levels: 0, 1 and 2.

[0013] Level 0 represents a common architecture present in modern computer systems, (see FIG. 1 and buses 120 and devices 110), where multiple buses 280 are coupled with multiple I/O devices 290. Next, level 1, also a common architecture present in modern computer systems, (see FIG. 1 bus-bridge 130), contains multiple bus-bridges 260 connected via grandchild-links 295 to multiple buses 280 where each such grandchild-link 295 has its own TOQs 270 for ordering the transactions for each individual bus 280. Also, each child-bridge 260 has one transaction buffer, i.e. 352, for each grandchild link, i.e., 296. Level 1's architecture, which provides a TOQ 270 for each connected bus 280, allows transaction ordering to occur in its most efficient form, i.e., one TOQ per one bus.

[0014] Finally, in FIG. 2, level 2 represents a disclosed embodiment which utilizes a bridge-bridge 232 with child-links 285 to the multiple bus-bridges 260 in level 1. Here, each child-link 285 to bus-bridges 260, has its own set of TOQs 320 and its own transaction buffer 349. The

number of TOQs 240 in any such set 320 is equal to the number of buses 280 attached via grand-child links 295 to the bus-bridge 260. In operation, when a new transaction is received by a transaction buffer, i.e., 349, the buffer contacts the appropriate TOQ 320, based on the grand-child link 295 the transaction originated. This one-to-one ratio of TOQs-to-buses at a bridge-bridge level (level 2), unlike simply a one-to-one ratio of TOQs-to-bus-bridge, allows pure transaction ordering, (i.e., one TOQ per bus), to occur in a bridge located at least one level deeper than the architecture shown in FIG 1. As such, FIG. 2's level 2 design represents a more efficient design than that shown in FIG. 1's level 2 design. Specifically, the potential blocking of any particular bus's transactions by another bus, as discussed above in relation to FIG. 1, is no longer present with the design disclosed in FIG. 2, i.e., where the disclosed multiple-TOQ concept is present.

[0015] In further detail, level 2's TOQs are broken into as many sets 320 as there are child-links 285. Four child-links 286, 287, 288 and 289 in parent-bridge 232 are shown, and each link has an associated TOQ set 320: 321, 322, 323 and 324 respectively, as well as their own transaction buffers 349. Further, within each TOQ set 320 are as many TOQs as there are grand-child links 295 for the associated child-link 285. For example, TOQ set 321, associated with child-link 286, and where such child-link 286 has four grand child links associated thereto: 296, 297, 298 and 299, is made up of four TOQs: 241, 242, 243 and 244. It should be noted that TOQ 243 is drawn in phantom form to show that it could represent multiple TOQs to assure that there were an equal amount of TOQs in set 321 as grand-child links associated with child-link 286. Each of the TOQ sets 320 contains a phantom TOQ for the same purpose. The remaining TOQ sets disclosed are as follows: TOQ set 322 contains TOQs 245, 246, 247 and 248; TOQ set 323, representing none or more TOQ sets 320, contains TOQs 249, 250, 251 and 252; and TOQ set 324 contains TOQs 253, 254, 255 and 256. Other embodiments may use less than one TOQ per grand-child link 295 for the associated child-link 285, but a minimum of two such TOQs are needed to optimize transaction ordering. Further, other multi-TOQ architectures may use more or less number of links to more or less number of child bridges.

[0016] As discussed above, each of the TOQs 240 in parent-bridge 232 correspond with a grandchild-link 295. As such, each TOQ 240 is matched with a grandchild-link 295. A matching can be achieved in variety of ways. A matching can occur where the child-bridge transmits or originates a transaction identifier that is a predefined address of an associated TOQ 240. For example, in the case of TOQ set 321 having four TOQs 241, 242, 243 and 244, TOQ

could have corresponding addresses 00, 01, 10 and 11. Thus, at the same time that child-bridge 260 is transmitting a transaction over child-link 286, the child-link 286 could transmit address 01 on a transaction identifier communication link, (two unused channels on child-link 286 for example), such that parent-bridge 232 routes the transmission to TOQ 242. It is also contemplated that rather than identifying a TOQ by an address, it may be identified with a stored key. Here, using the same transaction identifier 01, if each TOQ 240 had associated with it a key, the parent-bridge 232 could compare an incoming transaction identifier with each of the keys of each TOQ, and where there was a match, the parent-bridge would route the corresponding transmission to that TOQ, in this example again, TOQ 242. It is also contemplated that where a transaction is transmitted with a transaction identifier that does not match a key in any one of the TOQs 240, that the parent-bridge would route the transaction to a default TOQ, for example, TOQ 241. This assures that all transactions are handled.

[0017] In theory, the number of TOQs one could place within bridge-bridge 232 is limitless. However, in reality the number of TOQs employed in bridge-bridge 232 is limited by chip hardware constrains, such as size and complexity, and/or the ability and efficiency of uniquely identifying a transaction from a particular bus. In the disclosed embodiment, it is contemplated that a transaction is identified through a simultaneous transmission of an identifier from associated bus-bridge 260 to bridge-bridge 232 through an unused channel on the child-links 285. In the case of a bus-bridge 260 having two buses attached via grandchild-links 295, bus 281 and bus 282, for example, a single channel could be used to transmit an identifier of either a "0" or a "1" to indicate which bus was the source of the transaction. However, where more than two buses 280 are attached to the bus-bridge 260, and a constraint exists which requires the use of only one channel, or for which only two TOQs are available, it is contemplated that the same single channel could still be used to handle the transactions by assigning buses 280, or grand-child links 295, an identifier of either "1" or "0," where the separate TOQs would handle buses of common identities as though they originated from a single source. As the number of buses rises, or the number of available TOQs increase, the need for additional channels arise. For example, where 4 buses and 4 TOQs are present, two channels (i.e., base 10's "3" = base 2's "11") would be needed, if however 8 buses and 8 TOQs are present, three channels (i.e., base 10's "7" = base 2's "111") would be needed. It should be noted that other means of identifying a transaction may occur through signals sent through dedicated connections between the bus-bridge 260 and the bridge-bridge 232, or means other than an unused channel in child-links 285.

[0018] It is contemplated that if the disclosed embodiment is implemented with a typical parent-bridge 140 as found in FIG. 1, i.e., a parent-bridge with only one TOQ per child-links 285, rather than a parent-bridge 232 which has multiple TOQs per child-links 285, that this implementation would result in a system with the same throughput as the typical architecture shown in FIG. 1. This is because although a bus identification signal would be sent to the parent-bridge 232, there would be no functionality to receive it, nor any additional TOQs to take advantage of the information if it could be received, and thus, the parent-bridge 232 would simply order all the transactions coming through a particular child-link 285 as though they were originating from the same bus 280, or grand-child link 295, i.e., the same result as what is occurring at parent-bridge 140 in FIG. 1. Further, it is also contemplated that if the disclosed embodiment, having at least one parent-bridge 232 that in turn has multiple TOQs 240 per its child-links 286, is implemented without a child-bridge 260 that either originates or transmits bus identification signals to the parent-bridges 232, i.e, the child-bridge 160 of FIG. 1, that this implementation would also function with the same throughput as the architecture of FIG. 1. This is because the parent-bridge 232 would receive each transaction without any bus identification signal and for parent-bridge 232 to receive a transaction without an identification signal is the same as if it received a transaction with an address of "0." Thus, all the transactions received by parent-bridge 232 would be directed to a single TOQ resulting in the same throughput experienced by the system of FIG. 1.

[0019] FIGS. 3A and 3B is a disclosed embodiment that introduces the application of the disclosed techniques to architectures using three or more hierarchical levels of bridges. In the embodiment shown in FIGS. 3A and 3B, the same type of architecture and functionality that was attributed to the bridge-bridge 232 in level 2 in FIG. 2 (the parent-bridge in FIG. 2), is now in bridge-bridge 200 (parent-bridge) in level 3, but rather than ordering transactions received directly from a bus-bridge 260 in level 1, (the child-bridge in FIG. 2), the parent-bridge 200 receives transactions from a child-bridge 232 in level 2. Here, rather than TOQ sets 310 containing TOQs 218 that maintain ordering for buses 280, such TOQs 218 maintain ordering for bus-bridges 260 (grandchild-bridges). However, for embodiments that do not use the multi-TOQ design, the parent-bridge 200 in level 3, as shown in FIGS. 3A and 3B, receives multiple transactions over child-links 220 from the corresponding child-bridge 232, but does not know from which grandchild-bridge 260 the transaction originated, and therefore must apply standard transaction ordering across all transactions coming across child-links 220, thus, unnecessarily allowing the blocking of transactions originating from one grandchild-bridge 260 by transactions

originating from another grandchild-bridge 260. Regardless of the number of levels of hierarchical bridge architecture employed, the implementation of the multi-TOQ per grandchild-link design across any parent/child/grandchild combinations would result in improved throughput through such parent. Thus, architectures with three or more levels all have at least one parent/child/grandchild combination that can have their throughput optimized by including the suggested multi-TOQ architecture.

[0020] In further detail, level 3, in FIG. 3A, represents a disclosed embodiment which utilizes a bridge-bridge 200 with connections 220 (child-links) to the multiple bridge-bridges 230 disclosed in level 2. Here, every child-link 220 to bridge-bridges 260 has its own set of TOQs 310 and transaction buffers 339. The number of TOQs 218 in any such set 310 is equal to the number of bus-bridges 260 attached via grand-child links 285 to the bridge-bridge 230. This one-to-one ratio of TOQs-to-bus-bridges at a bridge-bridge level (level 3), allows separate transaction ordering to occur for the groups of transactions originating from any one bus-bridge, (i.e., one TOQ per bus-bridge), and to occur in a bridge located at least one level deeper than the architecture shown in FIG. 2. As such, FIG. 3's level 3 design represents a more efficient design than that shown in FIG. 2's level 2 design. Specifically, the potential blocking of any particular bus-bridge's transactions by another bus-bridge is no longer present with the design in FIGS. 3A and 3B, i.e., where the disclosed multiple-TOQ concept is introduced at a third level of bridges.

[0021] In even greater detail, level 3's TOQs are broken into as many sets 310 as there are child-links to bridge-bridge 200. Specifically, four child-links 221, 222, 223 and 224 are shown in parent-bridge 200, and each link has an associated TOQ set 310: 311, 312, 313 and 314 respectively, as well as their own transaction buffers 339. Further, within each TOQ set 310 there are as many TOQs 218 as there are grand-child links 285 for the associated child-link 220. For example, TOQ set 311, associated with child-link 221, and where such child-link 221 has four grand child links associated thereto: 286, 287, 288 and 289, is made up of four TOQs: 201, 202, 203 and 204. It should be noted that TOQ 203 is drawn in phantom form to show that it could represent multiple TOQs to assure that there were an equal amount of TOQs in set 311 as grand-child links associated with child-link 221. Each of the TOQ sets 310 contain a phantom TOQ for the same purpose. The remaining TOQ sets disclosed are as follows: TOQ set 312 contains TOQs 205, 206, 207 and 208; TOQ set 313 representing none or more TOQ sets 310, contains TOQs 209, 210, 211 and 212; and TOQ set 314 contains TOQs 213, 214, 215 and 216. Other embodiments may use less than one TOQ per grand-child link 285 for the associated child-

link 220, but a minimum of two such TOQs are needed to optimize transaction ordering. Further, other multi-TOQ architectures use more or less number of links to more or less number of child-bridges or grandchild-links.

[0022] Other embodiments may incorporate the disclosed multiple TOQ concept, but may do so in a fashion such that there is not a one-to-one correlation between the TOQs in the parent level bridge-bridge to the number of grand-child links. However, such embodiments would use at least two TOQs per child-link in any TOQ set. For example, looking at FIG. 2 where the parent level is 232, here if the TOQ set 321 were changed to have only two TOQs 241 and 242, and child-link 285 remained connected to bus-bridge 262 that continued to have the four grand-child links 296, 297, 298 and 299 to the four buses 281, 282, 283 and 284, then each TOQ 241 and 242, for example, could each handle the transactions from two of the buses. In such a design, although there are two buses for each TOQ, i.e., TOQ 241 for buses 281/282 and TOQ 242 for buses 283/284, and therefore the transactions from both buses are subject to the blocking of transactions from the other, this design still represents an improvement to a design where no multiple TOQs are used, as each bus, or grandchild-link, can only have its transaction blocked by one other bus (i.e., the other bus sharing the same TOQ), not three (i.e., not the three other grand-child links).

[0023] Additionally, rather than a single transaction buffer being used per child-link in a parent bridge, other embodiments may utilize a one-to-one ratio of transaction buffers to TOQs. For example, in FIG. 2, instead of there being one transaction buffer 342 per TOQ set 321, there would instead be four transaction buffers, (i.e., Buff 1a, Buff 2b, Buff 2c and Buff 2d), for each TOQ 241, 242, 243 and 244. As such, there are as many transaction buffers 349 as TOQs 240 (this multi-transaction buffer to multi-TOQ is not shown in any figure), and as many TOQs as there are grand-child links. When a transaction is received by the parent-bridge 232, the parent-bridge 232 places the transaction in the proper transaction buffers 342 (multiple buffers not shown), i.e. the transaction buffer 342 that corresponds with the grandchild-link 295 from which the transaction originated. It is contemplated that such an embodiment having a one-to-one ratio of transaction buffers to TOQs, might be easier to implement. However, it is also contemplated that such an embodiment would be more expensive from the standpoint of silicon used.

[0024] FIG. 4 shows the disclosed embodiment of FIG. 2 incorporated into a computer system 600. The computer system 600 includes CPU nodes 610, I/O nodes 630, and switch matrix 620.

CPU nodes 610 include the four nodes 611, 612, 613 and 614. I/O nodes 630 include four I/O nodes 601, 602, 603 and 604. A switch fabric 620 is connected to CPU nodes 610 and to I/O nodes 630. The embodiment disclosed in FIG. 2 is shown in FIG. 4 as I/O node 0. Like FIG. 2, I/O node 0 contains a parent-bridge 230. Parent-bridge 230 is attached via child-links 285 to child-bridges 260. The child-bridges 260, in turn, are connected via grand-child links 295 to buses 230. Finally, buses 230 are attached to I/O devices 290. It is contemplated, although not required, that some or all of the other I/O nodes would adopt a similar architecture shown in detail in I/O node 0.

[0025] The foregoing disclosure and description of the various embodiments are illustrative and explanatory thereof, and various changes in the nodes, buses, signals, components, circuit elements, circuit configurations, and signal connections, as well as in the details of the illustrated circuitry and construction and method of operation may be made without departing from the spirit and scope of the invention.